CLAIMS

What is claimed is:

1. A method of assigning a sample to a known or putative class, comprising the steps of:

a) determining a weighted vote for one of the classes for one or more informative genes in said sample in accordance with a model built with a weighted voting scheme, wherein the magnitude of each vote depends on the expression level of the gene in said sample and on the degree of correlation of the gene's expression with class distinction; and

b) summing the votes to determine the winning class and a prediction strength,

wherein said sample is assigned to the winning class if the prediction strength is greater than a prediction strength threshold.

2. The method of Claim 1, wherein the prediction strength is determined by:

$$(V_{win} - V_{lose}) / (V_{win} + V_{lose}),$$

wherein V_{win} and V_{lose} are the vote totals for the winning and losing classes, respectively.

3. The method of Claim 2, wherein the number of informative genes used in the weighted voting scheme is at least 50.

4. The method of Claim 3, wherein the known class is a known disease class.

5. The method of Claim 4, wherein the disease class is a cancer disease class.

6. The method of Claim 5, wherein the cancer disease class is Acute Lymphoblastic Leukemia (ALL) or Acute Myeloid Leukemia (AML).

7. The method of Claim 6, wherein the informative genes is selected from a group consisting of: C-myb, Proteasome iota, MB-1, Cyclin, Myosin light chain, Rb Ap48, SNF2, HkrT-1, E2A, Inducible protein, Dynein light chain, Topoisomerase II β, IRF2, TFIIEβ, Acyl-Coenzyme A, dehydrogenase, SNF2, ATPase, SRP9, MCM3, Deoxyhypusine synthase, Op 18, Rabaptin-5, Heterochromatin protein p25, IL-7 receptor, Adenosine deaminase, Fumarylacetoacetate, Zyxin, LTC4 synthase, LYN, HoxA9, CD33, Adipsin, Leptin receptor, Cystatin C, Proteoglycan 1, IL-8 precursor, Azurocidin, p62, CyP3, MCL1, ATPase, IL-8, Cathepsin D, Lectin, MAD-3, CD11c, Ebp72, Lysozyme, Properdin and Catalase.

8. The method of Claim 1, wherein the known class is a class of individuals who respond well to chemotherapy or a class of individuals who do not response well to chemotherapy.

9. A method of determining a weighted vote for an informative gene to be used in classifying a sample to be tested, comprising:

a) determining a weighted vote for one of the classes for one or more informative genes in said sample, wherein the magnitude of each vote depends on the expression level of the gene in said sample and on the degree of correlation of the gene's expression with class distinction; and

b) summing the votes to determine the winning class.

10. The method of Claim 9, wherein the weighted vote determined according to:

$$V_g=a_g(x_g-b_g),$$

wherein V_g is the weighted vote of the gene, g; a_g is the correlation between gene expression values and class distinction; $b_g=\mu_1(g)+\mu_2(g))/2$ which is the average of the mean $\log_{10}$ expression value in a first class and a second class; x_g is the $\log_{10}$ gene expression value in the sample to be tested; and wherein a positive V value indicates a vote for the first class, and a negative V value indicates a vote for the second class.

11. The method of Claim 10, wherein the vote for the first class is determined by obtaining a sum of the absolute values of the positive votes for the first class, and the vote for the second class is determined by obtaining a sum of the absolute values of the negative votes for the second class.

12. The method of Claim 11, wherein the weighted vote determined a portion of genes that are relevant for determining the classes.

13. The method of Claim 12, wherein a signal to noise routine, a Pearson correlation routine, or a Euclidean distance routine determines the relevant genes.

14. The method of Claim13, wherein the signal to noise routine is:

$$P(g,c)= (\mu_1(g)-\mu_2(g))/(\sigma_1(g) + \sigma_2(g)),$$

wherein g is the gene expression value; c is the class distinction, $\mu_1(g)$ is the mean of the expression levels for g for the first class; $\mu_2(g)$ is the mean of the expression levels for g for the second class; $\sigma_1(g)$ is the standard

deviation for the first class; and $\sigma_2(g)$ is the standard deviation for the second class.

15. A method for classifying a sample obtained from an individual into a class, comprising:

a) assessing the sample for a level of gene expression for at least one gene; and

b) using a model built with a weighted voting scheme, classifying the sample as a function of relative gene expression level of the sample with respect to that of the model.

16. The method of Claim 15, wherein assessing the level of gene expression comprises assessing the level of expression of a gene product.

17. The method of Claim 16, wherein the individual has a disease, and the sample is classified into a class of the disease.

18. The method of Claim 17, wherein the disease is cancer.

19. The method of Claim 18, wherein the cancer is leukemia.

20. The method of Claim 19, wherein the leukemia is AML or ALL.

21. A method for classifying a sample into a cancer disease class, wherein the sample is obtained from an individual and the level of gene expression for at least one gene is determined, comprising, using a model built with a weighted voting scheme, classifying the sample as a function of relative

gene expression level of the sample with respect to that of the model, to thereby classify the sample into the cancer disease class.

22. The method of Claim 4, wherein the cancer disease class is a leukemia class.

23. The method of Claim 5, wherein the leukemia class is AML or ALL.

24. A method for classifying a sample obtained from an individual, comprising:

a) subjecting the sample to at least one condition;

b) obtaining a gene expression product for two or more genes;

c) assessing the gene expression product for the genes to thereby determine the levels of the gene expression product for the genes;

d) using a computer model built with a weighted voting scheme, classifying the sample including comparing the gene expression levels of the sample to gene expression level of the model.

25. The method of Claim 24, wherein the genes assessed are the genes used to build the model.

26. In a computer system, a method for classifying at least one sample to be tested that is obtained from an individual, wherein gene expression values are determined for the sample to be tested, comprising:

a) receiving the gene expression values for the sample to be tested;

b) using a model built with a weighted voting scheme, classifying the sample including comparing the gene expression values of the sample to that of the model, to thereby produce a classification of the sample; and

c) providing an output indication of the classification.

27. The method of Claim 26, wherein the model is built according to:

$$V_g=a_g(x_g-b_g),$$

wherein V_g is the weighted vote of the gene, g; a_g is the correlation between gene expression values and class distinction; $b_g=\mu_1(g)+\mu_2(g))/2$ which is the average of the mean $\log_{10}$ expression value in a first class and a second class; x_g is the $\log_{10}$ gene expression value in the sample to be tested; and wherein a positive V value indicates a vote for the first class, and a negative V value indicates a negative vote for the class.

28. The method of Claim 27, wherein the vote for the first class is determined by obtaining a sum of the absolute values of the positive votes for the first class, and the vote for the second class is determined by obtaining a sum of the absolute values of the negative votes for the second class.

29. The method of Claim 28, wherein the weighted voting scheme builds the model using a portion of genes that are relevant for determining the classes.

30. The method of Claim 29, wherein a signal to noise routine, a Pearson correlation routine, or a Euclidean distance routine determines the relevant genes.

31. The method of Claim 30, wherein the signal to noise routine is:

$$P(g,c)= (\mu_1(g)-\mu_2(g))/(\sigma_1(g) + \sigma_2(g)),$$

wherein g is the gene expression value; c is the class distinction, $\mu_1(g)$ is the mean of the expression levels for g for the first class; $\mu_2(g)$ is the mean of the expression levels for g for the second class; $\sigma_1(g)$ is the standard deviation for the first class; and $\sigma_2(g)$ is the standard deviation for the second class.

32. In a computer system, a method for classifying at least one sample obtained from an individual, comprising:

a) providing a model built by a weighted voting scheme;

b) assessing the sample for the level of gene expression for at least one gene, to thereby obtain a gene expression value for each gene;

c) using the model built with a weighted voting scheme, classifying the sample comprising comparing the gene expression level of the sample to the model, to thereby obtain a classification; and

d) providing an output indication of the classification.

33. The method of Claim 32, wherein the model is built by a routine having:

$$V_g=a_g(x_g-b_g),$$

wherein V_g is the weighted vote of the gene, g; a_g is the correlation between gene expression values and class distinction; $b_g=\mu_1(g)+\mu_2(g))/2$ which is the average of the mean $\log_{10}$ expression value in a first class and a second class; x_g is the $\log_{10}$ gene expression value in the sample to be tested; and wherein a positive V value indicates a vote for the first class, and a negative V value indicates a negative vote for the class.

34. The method of Claim 33, wherein the vote for the first class is determined by obtaining a sum of the absolute values of the positive votes for the first class, and the vote for the second class is determined by obtaining a sum of the absolute values of the negative votes for the second class.

35. The method of Claim 34, wherein the weighted voting scheme builds the model using a portion of genes that are relevant for determining the classes.

36. The method of Claim 35, wherein a signal to noise routine, a Pearson correlation routine, or a Euclidean distance routine is used to determine the relevant genes.

37. The method of Claim 36, wherein the signal to noise routine is:

$$P(g,c)= (\mu_1(g)-\mu_2(g))/(\sigma_1(g) + \sigma_2(g)),$$

wherein g is the gene expression value; c is the class distinction, $\mu_1(g)$ is the mean of the expression levels for g for the first class; $\mu_2(g)$ is the mean of the expression levels for g for the second class; $\sigma_1(g)$ is the standard deviation for g the first class; and $\sigma_2(g)$ is the standard deviation for the second class.

38. In a computer system, a method for constructing a model for classifying at least one sample to be tested having a gene expression product, comprising:

a) receiving a vector for gene expression values of two or more samples belonging to more than one class, the vector being a series of gene expression values for the samples;

b) determining genes that are relevant for classification of a sample to be tested; and

c) using a weighted voting routine, constructing the model for classifying the samples using at least a portion of the genes determined in step B).

39. The method of Claim 38, wherein the step of determining employs a signal to noise routine, a Pearson correlation routine, or a Euclidean distance routine to determine the relevant genes.

40. The method of Claim 39, wherein the signal to noise routine is:

$$P(g,c)= (\mu_1(g)-\mu_2(g))/(\sigma_1(g) + \sigma_2(g)),$$

wherein g is the gene expression value; c is the class distinction, $\mu_1(g)$ is the mean of the expression levels for g for a first class; $\mu_2(g)$ is the mean of the expression levels for g for a second class; $\sigma_1(g)$ is the standard deviation for g the first class; and $\sigma_2(g)$ is the standard deviation for the second class.

41. The method of Claim 40, wherein the a weighted voting routine employs:

$$V_g=a_g(x_g-b_g),$$

wherein V_g is the weighted vote of the gene, g; a_g is the correlation between gene expression values and class distinction; $b_g =\mu_1(g)+\mu_2(g))/2$ which is the average of the mean $\log_{10}$ expression value in a first class and a second class; x_g is the $\log_{10}$ gene expression value in the sample to be tested; and wherein

a positive V value indicates a vote for the first class, and a negative V value indicates a negative vote for the class.

42. The method of Claim 41, wherein the vote for the first class is determined by obtaining a sum of the absolute values of the positive votes for the first class, and the vote for the second class is determined by obtaining a sum of the absolute values of the negative votes for the second class.

43. The method of Claim 42, further comprising performing cross-validation of the model.

44. The method of Claim 43, wherein performing cross-validation of the model comprises:

a) eliminating a sample used to build the model;

b) using a weighted voting routine, building a cross-validation model for classifying without the eliminated sample;

c) using the cross-validation model, classifying the eliminated sample including comparing the gene expression values of the eliminated sample to level of gene expression of the cross-validation model; and

d) determining a prediction strength of the class for the eliminated sample based on the cross-validation model classification of the eliminated sample.

45. The method of Claim 44, wherein the prediction strength is:

$$PS= (V_{win}-V_{lose})/V_{win}+V_{lose})$$

wherein V_{win} is the number of votes for the class to which the sample belongs, and V_{lose} the number of votes for the class to which the sample does not belong.

46. The method of Claim 38, further comprising filtering out any gene expression values in the sample that exhibit an insignificant change.

47. The method of Claim 38, further comprising normalizing the gene expression value of the vectors.

48. A computer apparatus for classifying a sample into a class, wherein the sample is obtained from an individual, wherein the apparatus comprises:

a) a source of gene expression values of the sample;

b) a processor routine executed by a digital processor, coupled to receive the gene expression values from the source, the processor routine determining classification of the sample by comparing the gene expression values of the sample to a model built with a weighted voting scheme; and

c) an output assembly, coupled to the digital processor, for providing an indication of the classification of the sample.

49. The computer apparatus of Claim 48, wherein the model is built according to:

$$V_g=a_g(x_g-b_g),$$

wherein V_g is the weighted vote of the gene, g; a_g is the correlation between gene expression values and class distinction; $b_g=\mu_1(g)+\mu_2(g))/2$ which is the average of the mean $\log_{10}$ expression value in a first class and a second class;

x_g is the $\log_{10}$ gene expression value in the sample to be tested; and wherein a positive V value indicates a vote for the first class, and a negative V value indicates a negative vote for the class.

50. The computer apparatus of Claim 49, wherein the vote for the first class is determined by obtaining a sum of the absolute values of the positive votes for the first class, and the vote for the second class is determined by obtaining a sum of the absolute values of the negative votes for the second class.

51. The computer apparatus of Claim 49, wherein the output assembly comprises a display of the classification.

52. A computer apparatus for constructing a model for classifying at least one sample to be tested having a gene expression product, wherein the apparatus comprises:

a) a source of vectors for gene expression values from two or more samples belonging to two or more classes, the vector being a series of gene expression values for the samples;

b) a processor routine executed by a digital processor, coupled to receive the gene expression values of the vectors from the source, the processor routine determining relevant genes for classifying the sample, and constructing the model with a portion of the relevant genes by utilizing a weighted voting scheme.

53. The computer apparatus of Claim 52, further comprising an output assembly, coupled to the digital processor, for providing the model.

54. The computer apparatus of Claim 52, wherein a weighted voting routine employs:

$$V_g=a_g(x_g-b_g),$$

wherein V_g is the weighted vote of the gene, g; a_g is the correlation between gene expression values and class distinction; $b_g=\mu_1(g)+\mu_2(g))/2$ which is the average of the mean $\log_{10}$ expression value in a first class and a second class; x_g is the $\log_{10}$ gene expression value in the sample to be tested; and wherein a positive V value indicates a vote for the first class, and a negative V value indicates a negative vote for the class.

55. The computer apparatus of Claim 54, wherein the vote for the first class is determined by obtaining a sum of the absolute values of the positive votes for the first class, and the vote for the second class is determined by obtaining a sum of the absolute values of the negative votes for the second class.

56. The computer apparatus of Claim 54, wherein the relevant genes are determined by a signal to noise routine, a Pearson correlation routine, or a Euclidean distance routine.

57. The computer apparatus of Claim 56, wherein the signal to noise routine is:

$$P(g,c)= (\mu_1(g)-\mu_2(g))/(\sigma_1(g) + \sigma_2(g)),$$

wherein g is the gene expression value; c is the class distinction; $\mu_1(g)$ is the mean of the expression levels for g for the first class; $\mu_2(g)$ is the mean of the expression levels for g for the second class; $\sigma_1(g)$ is the standard

deviation for g the first class; and $\sigma_2(g)$ is the standard deviation for the second class.

58. The computer apparatus of Claim52, further comprising a filter, coupled between the source and the processor routine, for filtering out any of the gene expression values in a sample that exhibit an insignificant change.

59. The computer apparatus of Claim 52, further comprising a normalizer, coupled to the filter, for normalizing the gene expression values.

60. The computer apparatus of Claim52, wherein the output assembly comprises a display of the model.

61. The computer apparatus of Claim 60, wherein the output assembly comprises a graphical representation.

62. The computer apparatus of Claim 61, wherein the graphical representation is color coordinated.

63. The computer apparatus of Claim 62, wherein the color coordination comprises shades of contiguous colors.

64. A machine readable computer assembly for classifying a sample into a class, wherein the sample is obtained from an individual, wherein the computer assembly comprises:

a) a source of gene expression values of the sample;

b) a processor routine executed by a digital processor, coupled to receive the gene expression values from the source, the processor

routine determining classification of the sample by comparing the gene expression values of the sample to a model built with a weighted voting scheme; and

c) an output assembly, coupled to the digital processor, for providing an indication of the classification of the sample.

65. The method of Claim 5, wherein the cancer disease class is glioblastoma or medulloblastoma.

66. The method of Claim 5, wherein the cancer disease class is folicular lymphoma or diffuse large B cell lymphoma.

67. The method of Claim 18, wherein the cancer is a brain tumor.

68. The method of Claim 67, wherein the brain tumor is medulloblastoma or glioblastoma.

69. The method of Claim 18, wherein the cancer is Non-Hodgkin's lymphoma.

70. The method of Claim 69, wherein the lymphoma is folicular lymphoma or diffuse large B cell lymphoma.